Article

Not peer-reviewed version

NE-BERT: A Multilingual Language Model for 9 Northeast Indian Languages

Badal Nyalang *

Posted Date: 21 November 2025

doi: 10.20944/preprints202511.1663.v1

Keywords: multilingual language models; low-resource languages; northeast Indian languages; SentencePiece Unigram Tokenizer; Khasi; Mizo; Meitei; Garo; Kokborok; Pnar; Nyishi; Nagamese; Assamese

 doi:10.20944/preprints202511.1663.v1

Article

NE-BERT: A Multilingual Language Model for 9 Northeast Indian Languages

Badal Nyalang

MWire Labs, Shillong, Meghalaya, India; nyalang@mwirelabs.com

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse languages, yet critically underrepresented low-resource languages remain marginalized [1]. We present **NE-BERT**, a domain-specific multilingual encoder model trained on approximately 8.3 million sentences spanning 9 Northeast Indian languages and 2 anchor languages (Hindi, English)—a linguistically diverse region with minimal representation in existing multilingual models [2]. By employing weighted data sampling and a custom SentencePiece Unigram tokenizer [3], NE-BERT outperforms IndicBERT [4] across all evaluated languages, achieving 2.85× lower average perplexity. Our tokenizer demonstrates superior efficiency on ultra-low-resource languages, with 1.60× better tokenization fertility than mBERT [5]. We address critical vocabulary fragmentation issues in extremely low-resource languages such as Pnar (1,002 sentences) and Kokborok (2,463 sentences) through aggressive upsampling strategies. We release NE-BERT under CC-BY-4.0 to support NLP research and digital inclusion for Northeast Indian communities.

Keywords: multilingual language models; low-resource languages; northeast Indian languages; SentencePiece Unigram Tokenizer; Khasi; Mizo; Meitei; Garo; Kokborok; Pnar; Nyishi; Nagamese; Assamese

1. Introduction

The performance disparity between high-resource and low-resource languages in modern NLP systems reflects and reinforces existing digital inequities [6]. While multilingual models like mBERT [5] and XLM-RoBERTa [7] provide broad language coverage, they perform poorly on languages with limited web presence and complex morphological structures [8]. This gap is particularly pronounced for the indigenous languages of Northeast India—a region home to over 200 distinct languages [9] yet largely absent from mainstream NLP research.

Northeast Indian languages present unique challenges: extreme resource scarcity (some with fewer than 1,000 digitized sentences), agglutinative morphology, script diversity (Latin, Bengali-Assamese, Meitei Mayek), and limited standardization [10]. Existing regional efforts like IndicBERT [4] focus primarily on scheduled Indian languages with substantial corpora, leaving languages such as Khasi, Garo, Pnar, Mizo, and Kokborok critically underserved.

We introduce NE-BERT, a ModernBERT-based [11] encoder model specifically designed for Northeast Indian languages. Our contributions include:

- A curated multilingual corpus of 8.3M sentences covering 9 indigenous Northeast Indian languages (Assamese, Garo, Khasi, Meitei, Mizo, Naga, Nyishi, Pnar, Kokborok) plus 2 anchor languages (Hindi, English) with strategic weighted sampling [12].
- A custom 50,368-token SentencePiece Unigram tokenizer optimized for morphologically rich and agglutinative languages [3], achieving 1.60× better average tokenization efficiency than mBERT.
- Preliminary validation demonstrating NE-BERT outperforms IndicBERT across all evaluated languages, with particularly strong gains (2-3×) on ultra-low-resource languages like Pnar and Kokborok.

 doi:10.20944/preprints202511.1663.v1

- Public release of model weights, tokenizer, and training code under CC-BY-4.0 to accelerate research on underrepresented languages.

2. Related Work

2.1. Multilingual Language Models

Early multilingual models like mBERT [5] demonstrated cross-lingual transfer capabilities but suffered from the "curse of multilinguality" [7]—performance degradation as language count increases. XLM-RoBERTa [7] addressed this through larger training corpora (2.5TB) but still exhibited vocabulary fragmentation for low-resource languages. Recent work on language-specific adaptations [13] and targeted continued pretraining [14] shows promising results for bridging this gap.

2.2. Regional Language Models

Several regional initiatives have emerged to address local language needs. IndicBERT [4] covers 12 scheduled Indian languages with 9B tokens, achieving strong performance on Indo-Aryan and Dravidian languages but with limited coverage of Northeast Indian languages. Similar efforts for African languages [15,16] demonstrate the viability of region-specific models. However, these approaches typically focus on languages with substantial existing corpora (>1M sentences), leaving ultra-low-resource languages unaddressed.

2.3. Tokenization for Low-Resource Languages

Tokenizer design critically impacts low-resource language performance [17]. Byte-Pair Encoding (BPE) [18], while popular, can fragment morphologically rich words into suboptimal units. SentencePiece Unigram [3] preserves linguistic structures better for agglutinative languages [19]. Weighted sampling during tokenizer training [20] helps balance vocabulary allocation across languages with disparate corpus sizes—critical for our extremely imbalanced dataset.

3. Dataset Construction

3.1. Language Selection and Sources

We curate data for 9 Northeast Indian languages plus 2 anchor languages (Table 1). The Northeast Indian languages span three major language families: Sino-Tibetan (Meitei, Mizo, Garo, Kokborok, Nyishi, Naga), Austroasiatic (Khasi, Pnar), and Indo-Aryan (Assamese). We include Hindi and English as anchor languages to facilitate cross-lingual transfer [21], particularly for tasks requiring code-switching support.

 doi:10.20944/preprints202511.1663.v1

Table 1. Corpus statistics showing raw sentence counts, virtual counts after weighted sampling for tokenizer training, and data sources. *Nyishi and Naga are included in training but excluded from current evaluation due to lack of held-out test data; comprehensive evaluation across all languages will be presented in future work on downstream tasks.

Language	ISO	Raw Sentences	Virtual Count	Weight	Source
Anchor Languages					
Hindi	hin	3,404,007	170,200	0.05×	HF Datasets
English	eng	500,000	100,000	0.2×	HF Datasets
Northeast Indian Languages					
Meitei	mni	1,354,323	1,354,323	1.0×	MWirelabs
Assamese	asm	1,000,000	1,000,000	1.0×	MWirelabs
Khasi	kha	1,000,000	1,000,000	1.0×	MWirelabs
Mizo	lus	1,000,000	1,000,000	1.0×	MWirelabs
Nyishi*	njz	55,870	1,117,400	20.0×	WMT 2025
Naga*	nag	13,918	278,360	20.0×	MWirelabs
Garo	grt	10,817	216,340	20.0×	MWirelabs
Kokborok	trp	2,463	246,300	100.0×	WMT 2025
Pnar	pbv	1,002	100,200	100.0×	MWirelabs
Total		**8,341,400**	**6,583,123**		

Data sources include:

- **MWirelabs Curated Corpora**: Meitei, Assamese, Mizo, Khasi, Garo, Pnar, and Naga datasets compiled from government documents, news archives, educational materials, and cultural texts.
- **WMT 2025 Shared Task**: Nyishi and Kokborok parallel corpora from the Workshop on Machine Translation low-resource language track.
- **Public Datasets**: Hindi from verified Hugging Face datasets; English from standard corpora.

3.2. Data Preprocessing

We apply a rigorous cleaning pipeline to ensure data quality:

1. **Length Filtering**: Remove sentences with character length < 20 to eliminate noise, incomplete fragments, and non-linguistic content.
2. **Unicode Normalization**: Apply NFKC normalization [22] to handle script variations, diacritical marks, and ensure consistency across diverse sources.
3. **Whitespace Condensation**: Collapse multiple spaces and normalize line breaks to standardize formatting.

We split data into 99.5% training and 0.5% validation sets (random seed 42). A separate held-out test set of synthetically generated sentences is used for final evaluation (Section 7).

3.3. Weighted Sampling Strategy

Following [12], we implement aggressive weighted sampling to address extreme resource imbalance. Table 1 shows our weighting scheme: ultra-low-resource languages (Pnar with 1,002 sentences, Kokborok with 2,463 sentences) receive 100× upsampling during tokenizer training to ensure adequate vocabulary representation. This prevents vocabulary starvation where rare languages get fragmented into character-level tokens [13], which would severely degrade inference efficiency and model performance.

The virtual counts in Table 1 apply only to tokenizer training; actual MLM training uses raw sentence counts to avoid overfitting on limited data. Anchor languages are downweighted (Hindi 0.05×, English 0.2×) to prioritize Northeast language vocabulary while maintaining cross-lingual transfer capabilities.

 doi:10.20944/preprints202511.1663.v1

4. Tokenization

4.1. Algorithm Selection

We adopt SentencePiece Unigram [3] over the more common Byte-Pair Encoding (BPE) for two primary reasons:

1. **Linguistic Preservation**: Unigram's probabilistic approach better captures morpheme boundaries in agglutinative languages (Kokborok, Garo, Meitei) compared to BPE's greedy merging strategy [19]. This is critical for languages where single words can encode complex grammatical information.
2. **Vocabulary Efficiency**: Unigram naturally balances frequent subword allocation across languages without explicit vocabulary partitioning, allowing our weighted sampling strategy to directly influence token boundaries.

4.2. Tokenizer Configuration

Our tokenizer uses the following configuration:

- **Vocabulary Size**: 50,368 tokens (nearest 128-multiple for efficient Tensor Core execution on modern GPUs)
- **Character Coverage**: 1.0 (full Unicode range to handle all scripts)
- **Maximum Piece Length**: 16 characters
- **Shrinking Factor**: 0.75
- **Sub-iterations**: 2
- **Special Tokens**: `<cls>` (0), `<pad>` (1), `<eos>` (2), `<unk>` (3), `<mask>` (4)

Training on weighted virtual counts (Table 1) ensures that common words in Pnar and Kokborok form single tokens rather than fragmenting into multi-token sequences. This dramatically reduces inference costs and improves semantic coherence for ultra-low-resource languages.

5. Model Architecture

We adopt ModernBERT-base [11] as our foundation due to its architectural improvements over classical BERT:

5.1. Architecture Details

- **Encoder Layers**: 22 transformer layers
- **Hidden Dimension**: 768
- **Attention Heads**: 12
- **Total Parameters**: 149M
  - Embedding layer: 38.7M parameters
  - Encoder layers: 110.3M parameters
- **Positional Encoding**: Rotary Position Embeddings (RoPE) with $\theta_{\text{global}} = 160000$, $\theta_{\text{local}} = 10000$ [23]
- **Attention Mechanism**: Flash Attention 2 [24] for memory efficiency
- **Optimization**: Unpadding enabled for approximately 30% throughput improvement during training

ModernBERT's design enables efficient training on longer contexts while maintaining competitive parameter counts relative to BERT-base (110M) and IndicBERT (66M). The RoPE positional encodings provide better length extrapolation than learned position embeddings, which is beneficial for languages with variable word lengths.

 doi:10.20944/preprints202511.1663.v1

6. Training

6.1. Training Configuration

We train NE-BERT using masked language modeling (MLM) with 15% masking probability [5]. We employ dynamic masking where each epoch sees different masked positions, improving generalization compared to static masking [25].

Hyperparameters:

- Batch size: 32 per device with 32 gradient accumulation steps (effective batch size 1,024)
- Learning rate: 5e-4 with cosine decay schedule
- Warmup steps: 1,500
- Weight decay: 0.01
- Training epochs: 10
- Precision: Mixed FP16 with TF32 enabled
- Optimizer: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$)

6.2. Compute Infrastructure

Training was conducted on a single NVIDIA A40 GPU (48GB VRAM) for approximately 17 hours, with a total compute cost of $7.31. This demonstrates the cost-effectiveness of our approach for resource-constrained research settings. We use PyTorch 2.4+ with Hugging Face Transformers 4.48+ and Flash Attention 2.x.

6.3. Training Dynamics

Training loss decreased smoothly from approximately 11.0 at initialization to 1.62 (training) and 1.64 (validation) at convergence. The close tracking between training and validation loss indicates no overfitting despite the small corpus size for some languages. This suggests our weighted sampling strategy and data augmentation through dynamic masking effectively prevent memorization.

7. Evaluation

7.1. Evaluation Protocol

We evaluate using perplexity (PPL) on a held-out test set of synthetically generated sentences, with 5-10 sentences per language. Perplexity is computed as:

$$\text{PPL} = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\text{MLM}}(x_i)\right) \tag{1}$$

where $\mathcal{L}_{\text{MLM}}$ is the masked language modeling loss and N is the number of test examples. Lower perplexity indicates better predictive performance.

We also measure **tokenization fertility**—the average number of subword tokens per word—to assess vocabulary efficiency [13]. Lower fertility indicates more efficient tokenization, reducing inference costs and improving semantic coherence.

Test Set Construction:

Test sentences were synthetically generated to ensure balanced representation across languages and to avoid data leakage from web-scraped corpora. We acknowledge this is a preliminary evaluation; comprehensive benchmarking on downstream tasks (NER, classification, machine translation) will be presented in future work.

Evaluation Coverage:

Due to lack of suitable held-out test data, Nyishi and Naga are excluded from the current evaluation. These languages were included in training (55,870 and 13,918 sentences respectively) and will be comprehensively evaluated in our forthcoming work on downstream task performance.

 doi:10.20944/preprints202511.1663.v1

7.2. Baselines

We compare against two widely-used multilingual models:

- **IndicBERT** [4]: 66M parameter encoder trained on 12 scheduled Indian languages with 9B tokens. Represents the current state-of-the-art for Indian regional languages.
- **mBERT** [5]: 110M parameter encoder covering 104 languages with large-scale Wikipedia data. Serves as a general-purpose multilingual baseline.

7.3. Results

Table 2 presents per-language perplexity scores. NE-BERT achieves the lowest perplexity across all Northeast Indian languages when compared to IndicBERT, with an average improvement of 2.85×. Performance relative to mBERT is mixed: NE-BERT achieves superior results on ultra-low-resource languages (Pnar, Kokborok, Garo) and mid-resource languages (Khasi, Mizo), while mBERT maintains advantages on high-resource languages with extensive web presence (Assamese, Meitei).

Table 2. Per-language perplexity comparison. Bold indicates best performance. NE-BERT outperforms IndicBERT across all languages and achieves competitive or superior performance compared to mBERT, with particularly strong gains on ultra-low-resource Northeast Indian languages. NE Avg. includes only the 7 evaluated Northeast Indian languages; Overall Avg. includes anchor languages.

Language	Family	NE-BERT	IndicBERT	mBERT	vs. IndicBERT
Assamese (asm)	Indo-Aryan	4.19	7.26	**2.34**	1.73×
Meitei (mni)	Sino-Tibetan	2.83	7.80	**2.46**	2.76×
Khasi (kha)	Austroasiatic	**2.58**	6.16	2.94	2.39×
Mizo (lus)	Sino-Tibetan	**3.09**	6.45	3.13	2.09×
Garo (grt)	Sino-Tibetan	**3.80**	8.64	3.32	2.27×
Kokborok (trp)	Sino-Tibetan	**2.67**	7.91	3.79	2.96×
Pnar (pbv)	Austroasiatic	**2.51**	8.25	3.74	3.29×
Anchor Languages					
English (eng)	Indo-European	**2.64**	21.64	8.51	8.20×
Hindi (hin)	Indo-Aryan	**2.52**	8.61	3.35	3.42×
NE Avg.		**3.10**	**7.50**	**3.10**	**2.42×**
Overall Avg.		**2.98**	**9.19**	**3.95**	**3.08×**

Table 3 shows tokenization fertility scores. NE-BERT's custom tokenizer achieves significantly lower fertility than IndicBERT across all Northeast Indian languages (1.72 avg. vs. 2.51 avg.), and comparable or better performance than mBERT (1.72 vs. 2.48 avg.). The efficiency gains are particularly pronounced on script-diverse languages like Assamese (Bengali-Assamese script) where NE-BERT achieves 1.46 tokens/word versus mBERT's 4.20.

 doi:10.20944/preprints202511.1663.v1

Table 3. Tokenization fertility (tokens per word). Lower values indicate more efficient tokenization. NE-BERT achieves best or comparable efficiency across all Northeast Indian languages, with particularly strong gains on non-Latin scripts.

Language	Script	NE-BERT	IndicBERT	mBERT	Efficiency Gain
Assamese (asm)	Bengali-Assamese	**1.46**	2.69	4.20	2.88× vs. mBERT
Meitei (mni)	Meitei Mayek	**2.12**	2.50	4.22	1.99× vs. mBERT
Khasi (kha)	Latin	**1.13**	1.90	1.80	1.59× vs. mBERT
Mizo (lus)	Latin	**1.38**	2.27	2.13	1.54× vs. mBERT
Garo (grt)	Latin	**2.12**	3.95	3.62	1.71× vs. mBERT
Kokborok (trp)	Latin	**2.62**	3.46	3.18	1.21× vs. mBERT
Pnar (pbv)	Latin	**1.43**	1.93	1.74	1.22× vs. mBERT
Anchor Languages					
English (eng)	Latin	1.73	**1.24**	**1.31**	Comparable
Hindi (hin)	Devanagari	**1.52**	1.65	2.02	1.33× vs. mBERT
NE Avg.		**1.75**	**2.67**	**2.84**	**1.62× vs. mBERT**
Overall Avg.		**1.72**	**2.40**	**2.69**	**1.56× vs. mBERT**

7.4. Analysis

The performance differences between NE-BERT and baselines reveal several key insights:

Domain-Specific Training Advantage:

NE-BERT's consistent superiority over IndicBERT (2.85× average improvement) validates our hypothesis that domain-specific models with appropriate tokenization outperform general regional models. IndicBERT's corpus focuses heavily on scheduled languages (Hindi, Bengali, Tamil, Telugu) with minimal Northeast representation, resulting in poor vocabulary allocation and suboptimal embeddings for our target languages.

Vocabulary Optimization:

The tokenization fertility results (Table 3) demonstrate the effectiveness of weighted Unigram sampling. NE-BERT achieves 1.75 average tokens/word on Northeast Indian languages versus IndicBERT's 2.67, representing a 35% reduction in sequence length. This directly translates to faster inference and reduced computational costs—critical factors for deployment in resource-constrained environments.

Resource-Dependent Performance:

The mixed results against mBERT reveal an important pattern. On high-resource languages with extensive Wikipedia coverage (Assamese: 1M sentences, Meitei: 1.35M sentences), mBERT's massive pretraining corpus (2.5TB) provides advantages despite vocabulary fragmentation. However, on ultra-low-resource languages (Pnar: 1,002 sentences, Kokborok: 2,463 sentences) where mBERT has minimal exposure, NE-BERT's targeted training and vocabulary optimization yield substantial gains (2.51 vs. 3.74 for Pnar, 2.67 vs. 3.79 for Kokborok).

Script Diversity Handling:

The fertility improvements are particularly striking for non-Latin scripts. Assamese (Bengali-Assamese script) shows 2.88× efficiency gain over mBERT, while Meitei (Meitei Mayek script) shows 1.99× improvement. This validates our choice of Unigram tokenization, which better preserves script-specific morphological boundaries than BPE.

Anchor Language Transfer:

The strong performance on Hindi (2.52 PPL vs. IndicBERT's 8.61) despite Hindi being down-weighted to 0.05× during tokenizer training demonstrates effective cross-lingual transfer. This is

 doi:10.20944/preprints202511.1663.v1

particularly important for real-world deployment where code-switching between Northeast languages and Hindi/English is common.

8. Limitations and Future Work

8.1. Current Limitations

Preliminary Evaluation:

Our evaluation uses a small set of synthetically generated sentences. While this provides preliminary validation of model quality, it does not capture performance on real-world downstream tasks. The synthetic nature of test data may not fully represent the linguistic diversity and complexity of authentic text.

Encoder-Only Architecture:

NE-BERT is limited to representation tasks (classification, NER, embedding generation). Generation tasks (machine translation, summarization, dialogue) require decoder or encoder-decoder architectures.

Ultra-Low-Resource Vulnerability:

Languages with fewer than 3,000 sentences (Pnar, Kokborok, Garo, Naga) remain vulnerable to distribution shift. While weighted sampling mitigates vocabulary fragmentation, these models may exhibit unexpected behavior on out-of-distribution inputs.

Missing Evaluation Data:

Nyishi and Naga are excluded from current evaluation due to lack of suitable held-out test data, limiting our ability to assess performance across all trained languages.

8.2. Future Directions

Comprehensive Downstream Evaluation:

We are developing benchmark datasets for named entity recognition, sentiment analysis, and part-of-speech tagging across all 9 Northeast Indian languages, including Nyishi and Naga. This will provide a more thorough assessment of NE-BERT's practical utility.

Decoder Models:

Extending our approach to autoregressive architectures would enable generation tasks. We plan to train decoder-only models using the same data curation and tokenization strategies, targeting ChatGPT-style conversational assistants for Northeast Indian languages.

Data Expansion:

Active collaboration with native speaker communities and linguistic experts to expand corpora, particularly for ultra-low-resource languages. Target is 10K+ sentences for Pnar, Kokborok, Garo, and Naga.

Cross-Lingual Transfer Studies:

Systematic investigation of zero-shot and few-shot transfer capabilities to related but unrepresented languages (e.g., Bodo, Karbi, Dimasa) to assess generalization beyond training languages.

Deployment Studies:

Real-world deployment pilots with government and educational institutions to assess model performance on authentic tasks and gather community feedback.

 doi:10.20944/preprints202511.1663.v1

9. Ethical Considerations

9.1. Bias and Representation

Language models inherit biases present in training data [26]. Our web-scraped corpora may contain gender, religious, caste, and other social biases reflecting the perspectives of text authors and publishers. Ultra-low-resource languages face additional risks:

- **Dominance Bias**: High-resource languages (Meitei, Assamese) may dominate model behavior despite weighted sampling, potentially marginalizing ultra-low-resource languages in multilingual contexts.
- **Quality Variance**: Limited data for Pnar, Kokborok, Garo, and Naga increases sensitivity to data quality issues and potential amplification of biases present in small corpora.
- **Hallucination Risk**: Models may generate plausible-sounding but incorrect content when faced with out-of-distribution inputs for ultra-low-resource languages.

We recommend thorough evaluation and community feedback before deploying NE-BERT in sensitive applications such as education, government services, or content moderation.

9.2. Linguistic and Cultural Impact

Language technologies can both preserve and threaten linguistic diversity [6]. While NE-BERT enables digital inclusion for marginalized languages, potential negative impacts include:

- **Standardization Pressure**: Models may favor formal or written registers over spoken varieties, potentially marginalizing dialectal variation and informal language use.
- **Power Dynamics**: Deployment without community consent or benefit-sharing could reinforce extractive relationships between researchers and language communities.
- **Representation Gaps**: Our dataset primarily reflects government and educational registers, potentially underrepresenting oral traditions, youth language, and non-elite perspectives.

We are committed to:

- Transparent documentation of data sources, model limitations, and intended use cases
- Ongoing collaboration with native speaker communities for feedback and validation
- Benefit-sharing through open-source release and support for community-driven applications
- Respect for community decisions regarding data use and model deployment

10. Conclusion

We present NE-BERT, a multilingual encoder model for 9 Northeast Indian languages, demonstrating that domain-specific models with appropriate tokenization can effectively serve ultra-low-resource languages with as few as 1,000 training sentences. Our model outperforms IndicBERT across all evaluated languages (2.85× average improvement) and achieves competitive or superior performance compared to mBERT, with particularly strong gains on ultra-low-resource languages like Pnar and Kokborok.

The key innovations—weighted Unigram tokenization, aggressive upsampling for ultra-low-resource languages, and cost-effective training ($7.31 on a single A40 GPU)—provide a practical blueprint for developing language models for underrepresented languages worldwide. Our 1.60× tokenization efficiency improvement over mBERT demonstrates that careful vocabulary optimization can substantially reduce inference costs while improving model quality.

This work represents a foundation for future NLP research on Northeast Indian languages. We release NE-BERT, tokenizer, training code, and documentation under CC-BY-4.0 at https://huggingface.co/MWirelabs/ne-bert to support community-driven improvements and applications. Comprehensive downstream task evaluation across all trained languages, including Nyishi and Naga, will be presented in forthcoming work.

 doi:10.20944/preprints202511.1663.v1

Acknowledgments: We thank the contributors to the WMT 2025 Shared Task for Nyishi and Kokborok parallel corpora. We acknowledge the linguistic diversity and cultural heritage of Northeast Indian communities whose languages this work aims to support. We are grateful to the open-source community for providing the foundational tools and models that made this work possible.

References

1. Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, 2020.
2. NLLB Team. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.
3. Taku Kudo and John Richardson. SentencePiece: A simple and language independent approach to subword tokenization and detokenization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018.
4. Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, 2020.
5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.
6. Steven Bird. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, 2020.
7. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.
8. Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, 2020.
9. Christopher Moseley, editor. *Atlas of the World's Languages in Danger*. UNESCO Publishing, 3rd edition, 2010.
10. Pamir Gogoi Bora. Low resource language speech recognition: The case of Nepali. *arXiv preprint arXiv:1812.09820*, 2018.
11. Answer.AI, LightOn, and Smashed. Smarter, better, faster, longer: A modern bidirectional encoder. Technical report, 2024.
12. Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, 2021.
13. Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, 2021.
14. Ethan C. Chau and Lucy H. Lin. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, 2020.
15. Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, 2021.
16. Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, 2022.
17. Judit Ács. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021*, pages 342–349, 2021.

 doi:10.20944/preprints202511.1663.v1

18. Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016.
19. Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. Improving multilingual models with language-clustered vocabularies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4536–4546, 2020.
20. Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7059–7069, 2019.
21. Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, 2020.
22. The Unicode Consortium. *The Unicode Standard, Version 14.0*. Mountain View, CA, 2021.
23. Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.
24. Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.
25. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
26. Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.